The University Of Sheffield.

Concepts of Statistics and their implementation for data analysis

Today's Outline

Part A : Statistics and its history
Populations and samples
Introduction of descriptive statistics

Break – 5 min

Part B : Frequencies and densities
Data plots
Introduction to probability and common distributions
Hypothesis testing
p-values

Class activity: Quiz –

Today we will be learning:

- To define populations and sample populations

- Accuracy, precision and Bias

- The meaning of descriptive statistics and their use in describing data

- The definition of frequencies and densities and their application in data analysis – general plots

- The definition of histograms and their use

- The concept of probabilities and definition of common distributions functions

- What are hypothesis testing and use of p-values

BMS353

What is Statistics?

Statistics is a science that studies the collection, the analysis, the interpretation, the presentation and organisation of data.

Based on mathematical tools, it allows to predict and forecast based on data that we observe (measure).

Statistics deals with all aspects of data including the planning of data collection in terms of the design of surveys and experiments. This means that we need to consider all aspects of the design and this starts with establishing an appropriate experimental design.

Adapted from Wikipedia

The Faces of Statistics

Sir William Petty, a 17th-century economist who used early statistical methods to analyse demographic data.

Thomas Bayes
1701-1761
Develop the Bayes' Theorem foundation for Bayesian inference

Pierre-Simon, marquis de Laplace, one of the main early developers of Bayesian statistics. 1749- 1827

Carl Friedrich Gauss, mathematician who developed the method of least squares in 1809. (1777-1855)

Karl Pearson, the founder of mathematical statistics. 1855-1936

Ronald Fisher, created the foundations for modern statistical science - 1890-1962

BMS353

Populations and Sample Populations

In statistics a *population* is the *total* set of observations that can be made of the group we are studying.

A population is any entire collection of people, animals, plants or things from which we observe data. It is the entire group under study that we wish to describe or draw conclusions about. We often need to study a representative sample of that population.

A *sample population* is a set consisting of observations that are taken from the population. The methods with which we draw these observation is called sampling.

For example, if we are studying the height of adult women in UK, the population is ALL the women in UK. If we are studying the average grade of students at Sheffield, the population is ….

Populations and Sample Populations (cont.)

The main difference between a population and sample there is to do with size, and how the observation are grouped:

A population includes **all** of the elements from a set of data.

A sample population consists of **one or more** observations from the population.

Sampling methods determine the *size* of the sample, that can have fewer observations than the population, the same number of observations, or more observations. More than one sample can be derived from the same population.

The population has quantifiable characteristics that are called *parameters* and denoted with Greek letters. Quantities like the mean or the standard deviations are parameters of the population.

Quantifiable characteristics of the sample population are called *statistics*. A sample statistic gives information about a corresponding population parameter. For example, the sample mean for a set of data would give information about the overall population mean ($\bar{x}$)

Populations and Sample Populations (cont.)

A *sampling method* is a procedure for selecting sample elements from a population. Simple random sampling refers to a sampling method that has the following properties.

- The population consists of N objects.
- The sample consists of n objects.
- All possible samples of n objects are equally likely to occur.

It is what we generally do all to describe population because enables us to use statistical analysis. If sampling is not random we cannot use it.

When a population element can be selected more than one time, we are *sampling with replacement*. When a population element can be selected only one time, we are *sampling without replacement*.

There are many different ways of random sampling, they are based on distribution like Normal, uniform, binomial etc..

Variables in statistics

Variables are things that we measure, control, or manipulate. They can be:
- Qualitative (names and lists) or Quantitative (counts and measurements).
- Continuous (can assume any value within a certain range) or discrete (defined over separate values, qualitative or quantitative)

Variables can be independent and not. Independent variables are those that are manipulated, whereas dependent variables are only measured or registered as a consequence of a manipulation on another variable.

Most commonly we measure variables independently and then look for relations (**correlations**) between them.

Descriptive statistics

Descriptive statistics is the term given to quantities that help to describe, **show or summarise** data in a meaningful way. They **DO NOT** enable to reach any conclusion on hypothesis that we might have made on the population.

They are important for: a) visualise the data; b) summarise the data

There are two general types of descriptive statistics:

Measures of central tendency: these are ways of describing the central position of a frequency distribution for a group of data. We can describe the central position using a number of statistics, including the mode, median, and mean.

Measures of spread: these are ways of summarizing a group of data by describing how spread out the data is. Measures of spread help us to summarize how spread out these scores are. We can describe the spread, using statistics including the range, quartiles, variance and standard deviation.

Descriptive statistics (cont.)

Measure of **central tendency** also known as measure of central **LOCATION**.

      mean

      median

      mode

Measure of **spread**, also known of measure of central **DISPERSION**.

      range

      variance

      standard deviation

      quartiles

Mean, Median and Mode

Mean: it is the "average" value: the sum of observations divided by the number of observations. The mean of a sample is an un-biased estimator of the population mean.

The sample mean is given by $\quad \bar{x} = (x_1 + x_2 + \ldots + x_n)/n$

And it is often represented by $\quad \bar{x} = \sum_{i=1}^{n} x_i / n \quad$ The population mean is the same $\quad \mu = \sum_{i=1}^{N} x_i / N$

Median: This is the value which divides the set of observations into two equal halves so that half the observations exceed the median and half are less than the median. Calculation of the median usually involves putting observations in RANK ORDER of size.

175 138 172 165 10 147 158

in rank order:

10 138 147 158 165 172 175

The median of this data is **158**.

Median or mean?
For highly skewed distributions, extreme untypical values can have a strong influence on the mean. In this particular instance, what does the observation "10" do to the mean?

The mean is **137.8**

BMS353

Mean, Median and Mode (cont.)

Mode: The mode is the value which occurs most frequently: the most "fashionable" value. A sample population can have more than one mode. For continuous or qualitative measurements, the **MODAL CLASS** is the class with the highest frequency ant it is nor necessarily unique.

examples
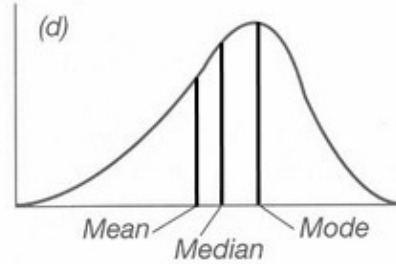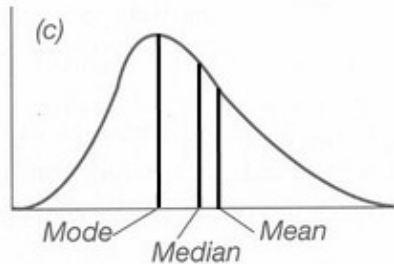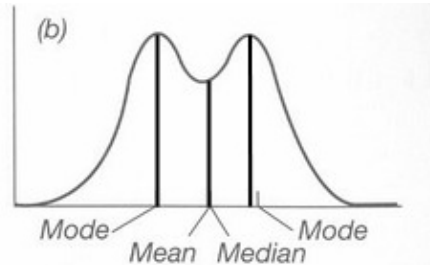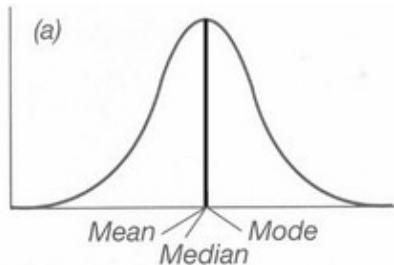
(a), (c) and (d) are unimodal
(c) is positively skewed
(d) Is negatively skewed

Type of Variable	Best measure
nominal	Mode
ordinal	Median
Interval/Ratio (not skewed)	Mean
Interval/Ratio (skewed)	Median

Range, Quartiles, Variance, and Standard Deviation

Range: This measures the **DIFFERENCE** between the largest and the smallest observation in the sample.

For example:
32.3 31.4 31.2 38.2 28.7 29.3 30.4 22.8 27.3 26.3

The largest observed value is <u>38.2</u> and the smallest is <u>22.8</u> hence
the **range** is **38.2-22.8 = 15.4**

The range is really easy to calculate, but there are a number of drawbacks with this measure:
It completely ignores the information provided by the remaining (n-2) observations and it is, by definition, strongly influenced by extreme untypical values.

Quartiles: are the three points that divide the data set into four equal groups, each group comprising a quarter of the data. The data is ranked before calculating the quartiles.

- The first quartile (Q1 or 25th percentile) is the middle number between the smallest number and the median of the data
- The second quartile (Q2 or 50th percentile) is the median of the data
- The third quartile (Q3 or the 75th percentile) is the middle value between the median and the highest value of the data
- IQR= interquartile range = Q_3-Q_1

Variance: This is one of the most important terms you will come across.
This measures the average squared deviation of observations from their **population** mean.

In mathematical terms, let us assume we are dealing with an entire population, consisting of *n* observations with population mean µ. Thus the variance of the population is

$$\sigma^2 = \sum_{i=1}^{N}(x_i - \mu)^2 / N$$

If all observations were exactly the population mean, then the average squared deviation of observations (hence variance) would be zero, but if they were more spread out then the variance would be higher. Why squared?

The variance is NOT an un-biased estimator of the population.

The variance within a population and the standard deviation within a population cannot be estimated directly from the variance and standard deviation calculated within a sample.
It is clear when we compare the range of a population with the range from a sample. Why?

The sample range will tend to underestimate population range. In the same way, the variance and standard deviation in a sample would tend to underestimate that of a population. To correct this, we use slightly altered formula to get un-biased estimators.

Variance of sample population $\qquad$ $\displaystyle var = \sum_{i=1}^{N} (x_i - \bar{x})^2 / (N-1)$

Standard deviation of the sample population $\qquad$ $\displaystyle sd = \sqrt{\sum_{i=1}^{N} (x_i - \bar{x})^2 / (N-1)}$

The **standard deviation** of an entire population is the square root of the variance of that population

Part B

Today's Outline

Part A : Statistics and its history
         Populations and samples
         Introduction of descriptive statistics

Break – 5 min (questions if any)

Part B : Frequencies and densities
         Data plots
         Introduction to probability and common distributions
         Hypothesis testing
         p-values

Class activity: Quiz – part A and B

Frequency distributions

A *frequency distribution* is a a table that displays the occurrence of various outcomes in a sample, each counted within a particular group or interval.

If you have a pool of 100 students and ask them to answer to the following question: *The use of statistics in biology is a great advantage*
You gather the following data
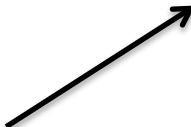
Strongly agree	20	Frequency or count
Agree somewhat	30	
Not sure	20	
Disagree somewhat	15	***Univariate frequency distribution***
Strongly disagree	15	

Frequency distributions (cont.)

You can also create ranges of values to group the data. For example if you are summarising the same pool of 100 students by their weight, you can gather the data in the following form:

	# of students	Cumulative #
Less then 50kg	25	25
50–55 kg	35	60
55–60 kg	20	80
60–65 kg	20	100

bins →

Cumulative distribution

Some of the graphs that can be used with frequency distributions are histograms, line charts, bar charts and pie charts. Frequency distributions are used for both qualitative and quantitative data.

Probability density function

A *probability density function* (PDF) or density of a continuous random variable x, is a function that describes the relative likelihood for this random variable to take on a given value. The probability density function is non-negative everywhere, and its integral over the entire space is equal to one.

$$P(a \leq x \leq b) = \int_a^b f_x(x)\,dx \qquad \text{or} \qquad P(X) = \int f_x(x)\,dx$$

The **sample space X** for a probability function is the set of all possible outcomes.
An *random variable representing an event* **x** is a subset of the sample space X.

A **probability** is a numerical value assigned to a given event x. The probability of an event is written *P(x)*, and describes the long-run relative frequency of the event. The first two basic rules of probability are the following:

> **Rule 1:** *Any probability P(x) is a number between 0 and 1 ($0 \leq P(x) \leq 1$).*

> **Rule 2:** *The probability of the sample space X is equal to 1 (P(X) = 1).*

Common probability distribution functions

We will look at the details of three probability distribution functions.

These are commonly used in data analysis and we will make extensive use of them in our module.

- Uniform distribution

- Binomial distribution *

- Normal distribution (Gaussian probability distribution)

The function in Week7 practical runif() and rnorm() allowed to build a sequence of values from uniform or normal distribution.
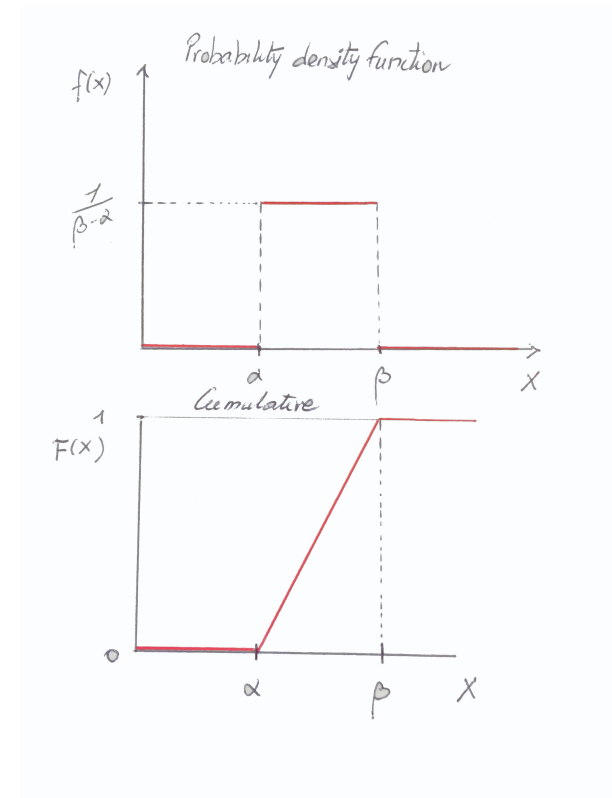
Uniform Distribution

Uniform distribution (also known as rectangular distribution) is a family of symmetric probability distributions that can be continuous or discrete

Continuous uniform distribution ($U(\alpha,\beta)$) is such that all intervals of the same length on the distribution's space are equally probable. The space is defined by the two parameters, α and β, which are its minimum and maximum values of the distribution.

$$f(x) = \begin{cases} \dfrac{1}{\beta - \alpha} & \text{for } \alpha \leq x \leq \beta \\ \\ 0 & \text{for } x < \alpha \text{ and } x > \beta \end{cases}$$

Later in the quiz… mean, median and mode?

Variance: $\sigma^2 = \dfrac{1}{12}(\beta - \alpha)^2$

Probability density function

Cumulative

Uniform Distribution (cont.)

It is when a finite number of values (N) are equally likely to be observed. Every values has equal probability 1/N to occur.

Probability mass function
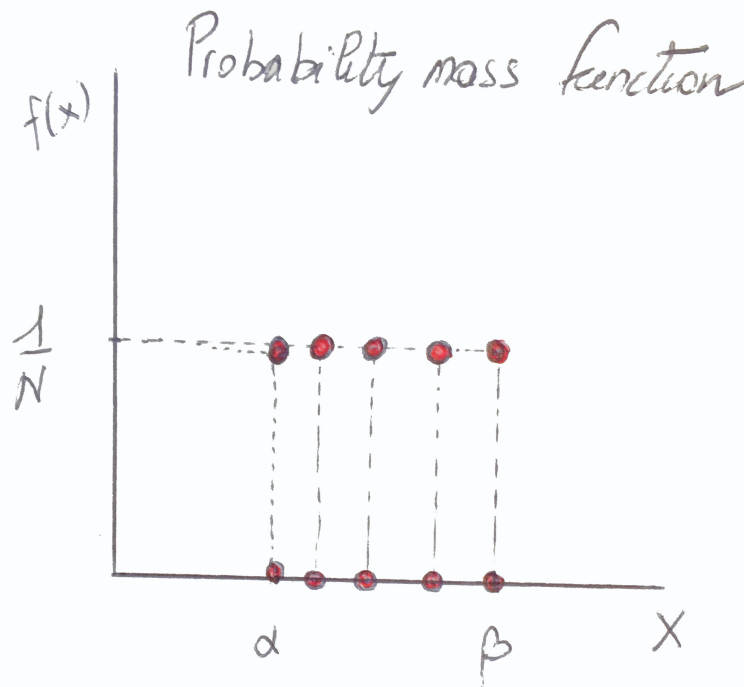
$F(x)=1/N$

Mean ?
Median ?
Mode ?

variance $\dfrac{(\beta-\alpha+1)^2-1}{12}$

Normal Distribution

The **normal distribution**, also called the **Gaussian distribution**, is an important family of continuous probability distributions, applicable in many fields. Each member of the family may be defined by two parameters, *location* and *scale*: **the mean (μ) and variance ($\sigma2$)** respectively. The **standard normal distribution** is the normal distribution with a mean of zero and a variance of one.

The importance of the normal distribution as a model of quantitative phenomena is due in part to the central limit theorem. Many measurements, ranging from psychological to physical phenomena (in particular, thermal noise) can be approximated, to varying degrees, by the normal distribution.
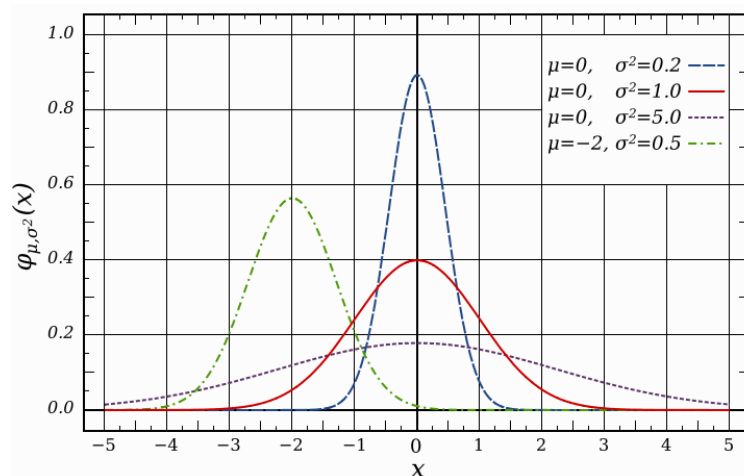
Probability density function:

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Mean = Median = Mode = μ
Variance = σ^2

Normal Distribution

One important property is that the **mean ± 1.96** standard deviations contain 95% of all observations while the **mean ± 2.576** standard deviations contain 99% of observations.
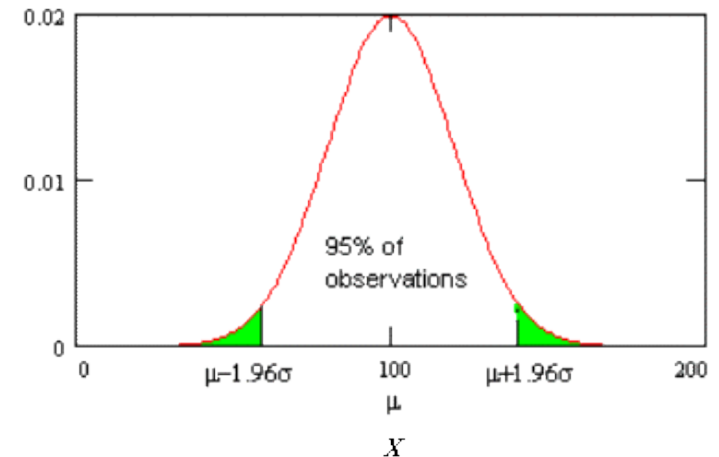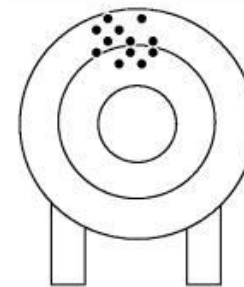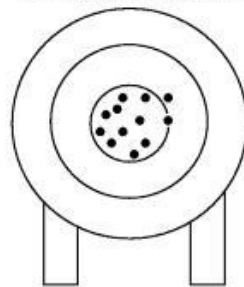
In R....
quantile(morley$Speed,prob=0.75)[["75%"]] + 1.5*IQR(morley$Speed)

Accuracy and Precision

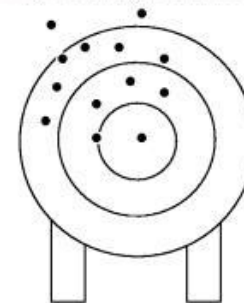**Accuracy** – how close the measurement are to the true value that your are trying to measure

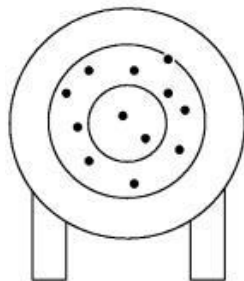**Precision** – how repeatable the measurements is (how fine resolution) respectable of whether is close to the actual value

Good accuracy
Good precision

Good precision
Poor accuracy

Poor Precision
Moderate accuracy

Poor accuracy
Poor Precision

Shooting target analogy

Bias

Bias – It is an error, a systematic lack of accuracy. The data Is deviating from the true value in the same direction.

It is therefore important to distinguish situations where the measurements differ from the true value at random and where they do systematically

Where is the bias?
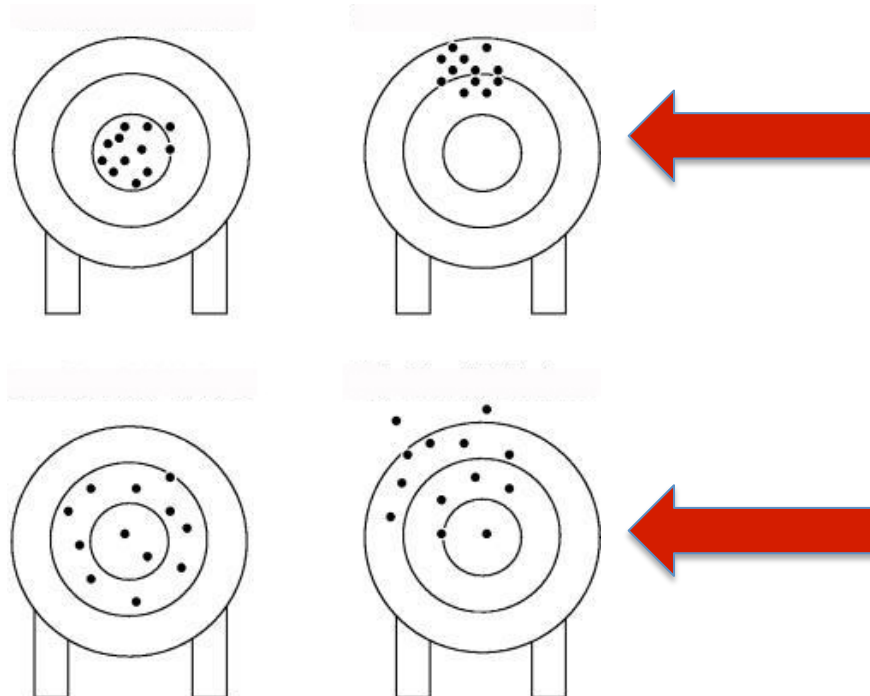
Data Visualisation: Bar Charts

Discrete data:

A way of summarising discrete (qualitative or quantitative) data is by counting the number of observations falling into each category.
The *relative frequency* is a number which describes the proportion of observations falling in a given category. We can plot the total frequency in a chart called **bar chart**
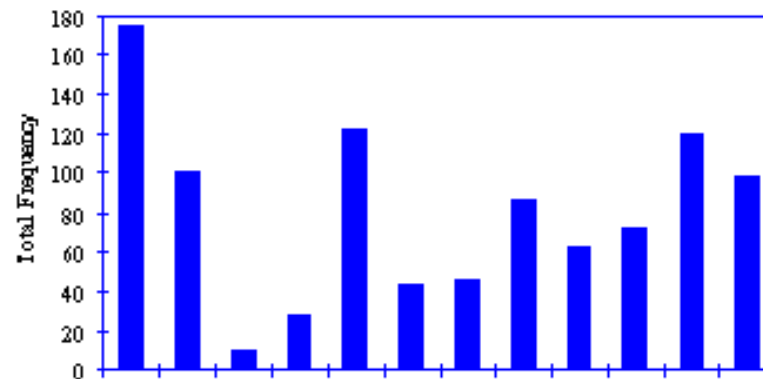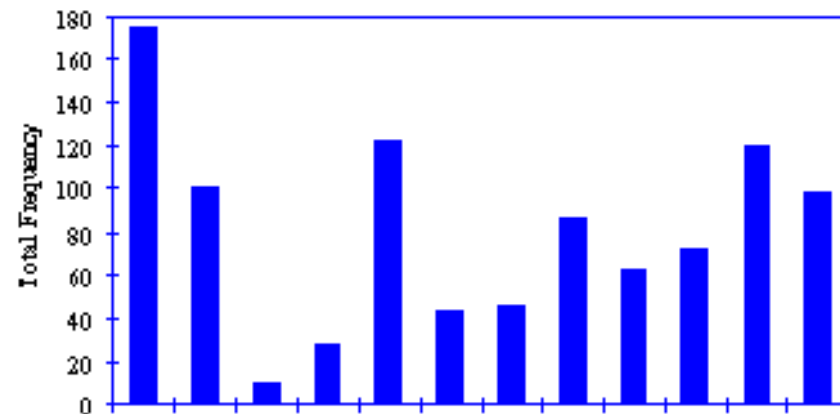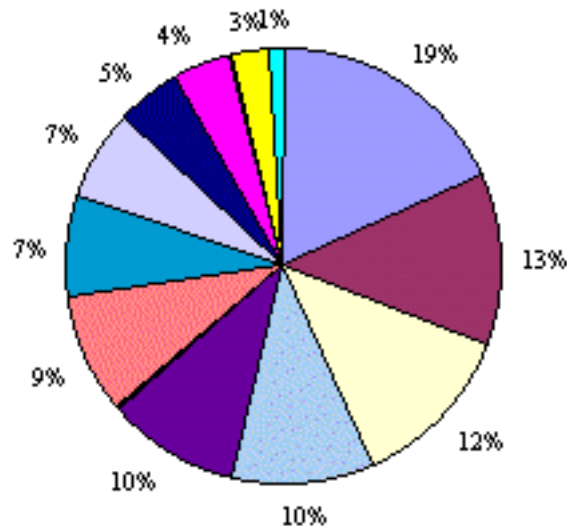

In R we use the command *barplot()*

barplot(game_cards$score, names = game_cards$names)

Data Visualisation: Pie Chart

The graphical representation of relative frequencies or percentages is a **Pie chart.**

It is more appropriate, when we need to highlight proportions rather than counts. Each slice of the pie represents a proportion of the total.

In R...
# Pie Chart from data frame with Sample Sizes
iris_table <- table(iris$Species)
lbls <- paste(names(iris_table), "\n", iris_table, sep="")
pie(iris_table, labels = lbls,
        main="Pie Chart of Species of Iris\n (sample sizes)")

Data visualisation: box plot

A **box plot** is a graphical representation of groups of numerical data through their quartiles. Box plots may also have lines extending vertically from the boxes (whiskers) indicating variability outside the upper and lower quartiles, hence the terms box-and-whisker plot and box-and-whisker diagram. Outliers may be plotted as individual points.
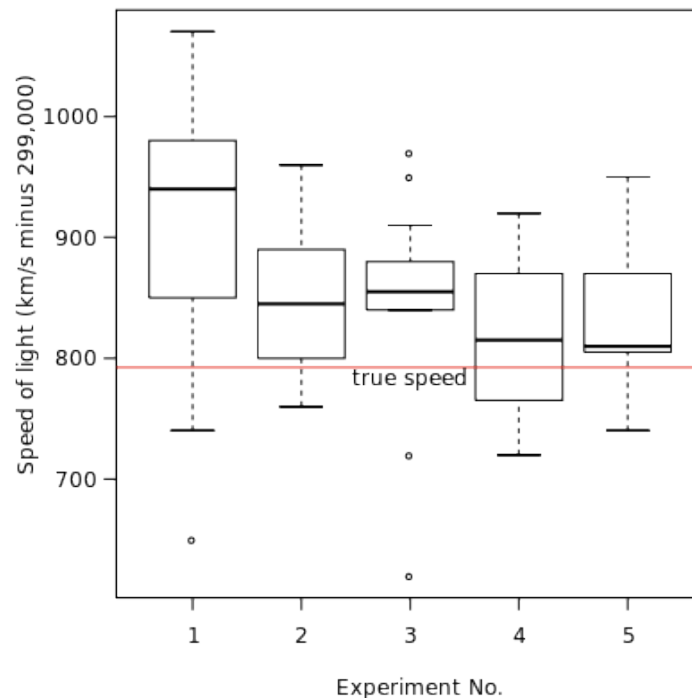
Box plot of data from the Michelson–Morley experiment, 1887

```
boxplot(morley$Speed ~ morley$Expt,
  col='light grey', xlab='Experiment #',
  ylab="speed (km/s - 299,000)",
  main="Michelson–Morley experiment")
mtext("speed of light data")
```

```
sol=299792.458-299000 # deviation of real speed of
ligth from the estimated 299,000 km/s
abline(h=sol, col='red')
```

Data visualisation: Histogram

It is an estimate of the *probability distribution* of a continuous variable (quantitative variable) and was first introduced by Karl Pearson

To estimate this **distribution** we can proceed in the same way as the bar chart but we first group the observations. This consists in choosing a set of contiguous non-overlapping intervals, called class intervals (or bins), the observations can be grouped to form a discrete variable from the continuous variable.

1. Identify suitable, non-overlapping classes in which to put this data. Generally, between 5 and 15 classes is best.

2. Count the number of observations that fall in the classes.
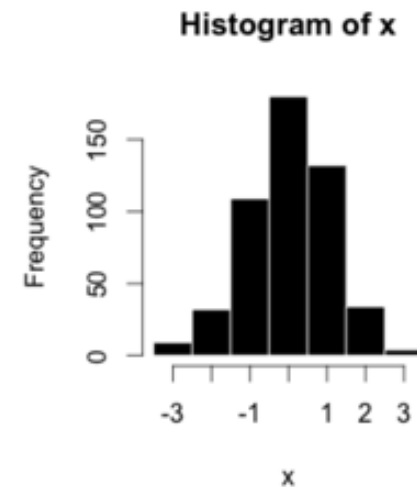
Histogram of x

Histograms give a rough sense of the *density* of the underlying distribution of the data and often are used as density estimators (to estimate *the probability density function* of the underlying variable)

Probability density function (cont.)
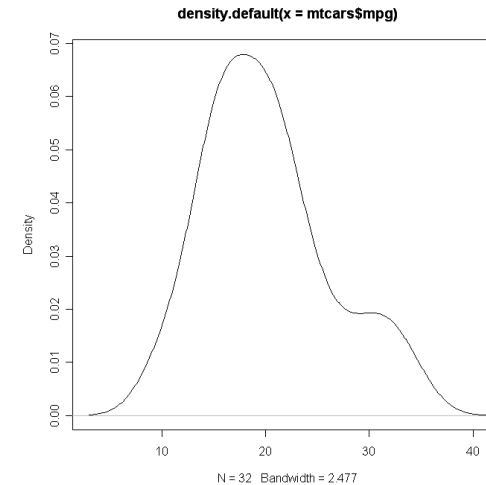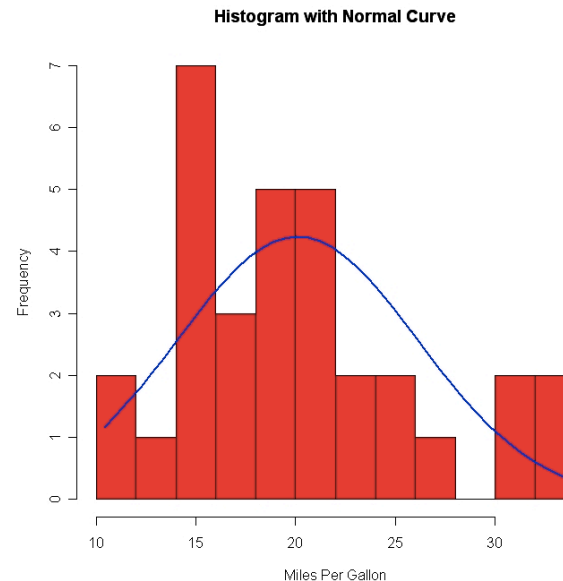

```
par(fg=rgb(0.6,0.6,0.6))
hist(morley$Speed, prob=F,
    col=rgb(0.9,0.9,0.9),
    main='Michelson-Morley Experiment ',
    ylab="Frequency", xlab='Difference from Speed of Light')
par(fg='black')

lines(density(morley$Speed))
```

Why is the histogram not estimating well the density function?

Statistical tests

The ultimate goal of most statistical tests is to evaluate relations between variables, in general to evaluate a ratio of some measure of the differentiation common in the variables in question to the overall differentiation of those variables.

When is this relationship significant?
The significance depends mostly on the **sample size**.
In very large samples, even very small relations between variables will be significant.
In very small samples even very large relations cannot be considered reliable (significant).

How do we quantify significance?
Thus, in order to determine the level of statistical significance, we need a function that represents the relationship between "magnitude" and "significance" of relations between two variables, depending on the sample size.

This function would give us the significance level (p-value), and it would tell us the probability of error involved in rejecting the idea that the relation in question does not exist in the population (null hypothesis).

Confidence and p-values

The statistical significance of a result is the probability that the observed relationship (e.g., between variables) or a difference (e.g., between means) in a sample occurred by pure chance, and that in the population from which the sample was drawn, no such relationship or differences exist.

In other words, statistical significance of a result tells us something about the degree to which the result is "true", is "representative of the population.

The measure that we normally use to quantify this is a **p-value**.
p-value represents a decreasing index of the reliability of a result. The higher the p-value, the less we can believe that the observed relation between variables in the sample is a reliable indicator of the relation between the respective variables in the population. Specifically, the **p-value represents the probability of error that is involved in rejecting the Null Hypothesis based on our observed data.**

Structure of a test

- Research question
  - Your investigative question
  - Ex: Do people have a preference for movie type?

- Hypotheses
  - The null hypothesis (H0), which is the hypothesis that states there is no significant difference between expected and observed data. Investigators either accept or reject H0.
  - **HO**: The observed distribution fits the expected or, in
    other words, there is no preference.
  - **HA**: The observed distribution does not fit that expected
    (there is a preference).

- Assumptions
  - The sample is chosen randomly.
  - The scores are independent (i.e., each subject is allowed only one preference).
  - The null hypothesis.

- Decision Rule
  - This is the threshold by which H0 is excepted or rejected – p-value threshold

- Computation
  - The calculation of the appropriate descriptive statistics

- Decision

BMS353

Type of tests

In the literal meaning of the terms, **a parametric statistical test** is one that makes assumptions about the parameters (defining properties) of the population distribution(s) from which the data are drawn. A **non-parametric test** is one that makes no such assumptions.

Nonparametric tests are also called distribution-free tests because they don't assume that the data follow a specific distribution.

When your data don't meet the assumptions of the parametric test, especially the assumption about being normally distributed, it is the case to use non-parametric tests. It is a basic rule , but there are additional considerations.
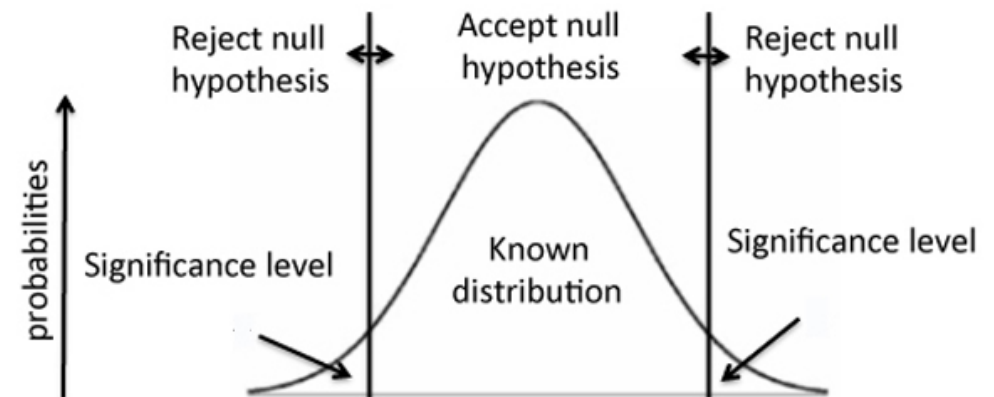
This data is described as **unpaired** or independent when the sets of data arise from separate individuals or **paired** when it arises from the same individual at different points in time.

One-tailed and two-tailed test

One-tailed test
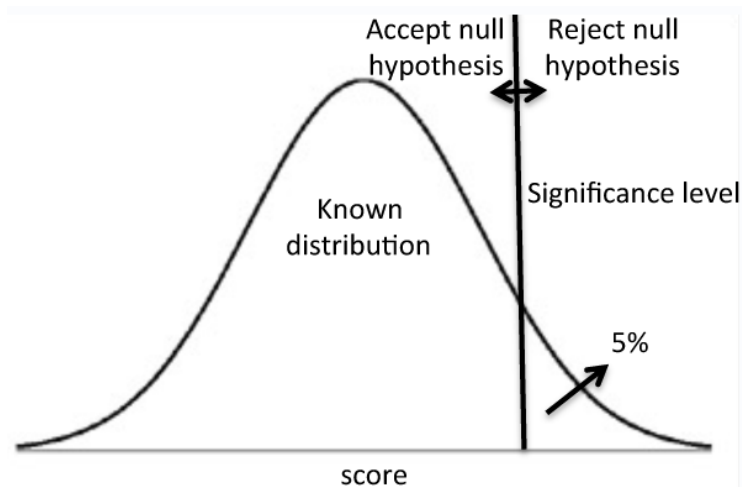
Two-tailed test

Significance level = p-value = τ = 0.05

$$t = (\bar{x} - \mu) / (s / \sqrt{n})$$ t score used in at test to compute the decision

- Known population normally distributed
- The sample is randomly selected
- The s of the unknown population is the same as the known population

Summary

- How to define populations and sample populations in your study

- Descriptive statistics and their use in data analysis

- Frequency and densities and their application in data analysis

- Probabilities and definition of common distributions functions

- Basic data representation tools and definition of histograms and their use

- What are hypothesis testing and use of p-values